基于距离和密度的 d-K-means 算法 *

唐泽坤，朱泽宇，杨　裔，李彩虹，李　廉
(兰州大学 信息科学与工程学院, 兰州 730000)

摘　要： K-means 算法有实现简单、速度快的特点，是应用最广泛的聚类算法。针对 K-means 算法对初始聚类中心和噪声敏感的缺点，提出了 d-K-means 算法（distance&density），在 K-means 算法的基础上权衡了密度和距离对聚类的影响，对数据进行加权处理，在权值基础上引入“最小最大原则”选择初始聚类中心，自动确定类中心个数。实验结果表明，d-K-means 算法在低维数据与高维数据上都可以取得较好的聚类效果，并且更好地应对低密度区域数据，更好地进行类中心选择。

关键词： 聚类；K-means 算法；最小最大原则；类中心个数
中图分类号： TP301.6　　doi: 10.19734/j.issn.1001-3695.2018.10.0861

D-K-means algorithm based on distance and density

Tang Zekun, Zhu Zeyu, Yang Yi, Li Caihong, Li Lian
(*College of Information Science & Engineering*, *Lanzhou University*, *Lanzhou* 730000, *China*)

Abstract: K-means algorithm has the characteristics of simple implementation and fast speed, and is the most widely used clustering algorithm. To deal with the disadvantages of the K-means algorithm which is sensitive to initial clustering center and noise, the d-K-means algorithm is proposed. On the basis of the K-means algorithm, the d-K-means algorithm weighs the impact of density and distance on clustering to weight the data, and selects the initial clustering center by introducing the "minimax principle" on the basis of weight, and automatically determines the number of class centers. Experimental results show that d-K-means algorithm can achieve better clustering results on low-dimensional data sets and high-dimensional data sets, and better deal with low-density regional data , and better select class centers.

Key words: clustering; K-means algorithm; minimax principle; number of class centers

0　引言

聚类是数据挖掘中将物理或抽象对象的集合分成由类似的对象组成的多个类的方法。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。“物以类聚，人以群分”，在自然科学和社会科学中，存在着大量的分类问题[1]。

K-means 算法因思想较为简单且易实现的特点，成为应用最广泛的聚类算法。但是 K-means 算法也有局限性，比如算法需要预先设定分类数量，用户在对数据了解不够的情况下难以给出合适的值；另一个局限性体现在算法初始中心设置的随机性使聚类结果易陷入局部最优解，并且聚类结果不稳定[2]。

近年来，针对以上缺陷，学术界对 K-means 算法[3,4]进行了多种优化研究。文献[5]计算所有样本点算术平均值组成初始聚类中心，考虑了全部数据样本的分布情况，改善了 K-means 算法初始中心点选择的随机性，但是易受噪声的影响。文献[6]构造最小生成树并利用树的减枝技术基于数据分布动态选取初始聚类中心，考虑了样本集密度对聚类的影响，但构造最小生成树较大程度上增加了算法开销。文献[5]和文献[6]都根据整个样本集数据分布选取初始中心点，忽略了真实聚类中心周围密度都较大，易陷入局部最优。文献[7]基于距离最远的样本点不可能分到同一个簇中，避免了陷入局部最优解的问题，但是没有考虑密度，可能选择噪声点作为初始类中心。文献[8]计算离群因子对数据集进行升序排序，使中心点位置靠前，考虑了密度的影响，结合“最小最大原则”选择初始中心点，不会陷入局部最优，但可能误将低密度区域数据归为离群点，影响最终聚类效果。文献[9]将数据对象根据密度排序，每次选取连接密度最大的数据点和与其距离最近的数据点的线段中点作为新的聚类中心，然后以此中心点为圆心，使指定半径的圆内数据点不参加接下来的聚类，一定程度上解决了文献[8]的问题；但是当低密度区域与高密度区域距离较大时，低密度区域数据依旧会被忽略。上述算法生成聚类中心上述算法都需要提前确定中心个数，在人们对数据集研究不透彻时会影响聚类结果精度。文献[10]根据密度参数产生初始聚类中心点候选集合，然后遍历集合中的数据进行模拟聚类，选择类间分离性和类内紧密性最优的数据点，充分考虑了最终聚类效果；但算法时间复杂度很高，密度参数的计算方法不科学，可能存在实际密度相差较大的数据点密度参数相同，可能导致高密度点在中心点选取过程中被移除，降低初始中心点质量。文献[11]将 canopy 算法思想与密度相结合，可以更好地处理低密度区

收稿日期： 2018-10-31；**修回日期：** 2019-01-07　**基金项目：** 国家重点研发计划资助项目；云计算和大数据专项项目；数据科学的若干基础理论项目（2018YFB1003205）；国家自然科学基金资助项目（61300230，61370219）；甘肃省自然科学基金资助项目（1107RJZA188）；甘肃省科技支撑计划资助项目（1104GKCA037）；甘肃省科技重大专项项目（1102FKDA010）

作者简介： 唐泽坤（1995-），男，硕士研究生，主要研究方向为资源服务、数据挖掘、数据预测(532460788@qq.com)；朱泽宇（1994-），男，硕士研究生，主要研究方向为数据挖掘、系统结构；杨裔（1980-），男，副教授，硕导，主要研究方向为机器学习、人工智能、语音识别；李彩虹（1973-），女，高级实验师，硕导，主要研究方向为模式识别、机器学习；李廉（1951-），男，教授，博导，主要研究方向为机器学习、计算机网络，无线传感器网络.

域数据，自动确定类中心个数，只有遍历完全部数据点后才会停止，没有考虑聚类效果，存在噪声点、离群点时很可能将其归为单独一类，影响聚类效果和类中心个数精度。

聚类中心点的选取原则是首先周围点相对较密集，其次聚类中心点之间距离相对较远，即具有很好的分布性，防止陷入局部最优解。为解决初始中心点选择和聚类中心数问题，本文提出基于距离和密度的 d-K-means 算法，采用加权的方法权衡密度与距离的关系，在全局范围求解，使中心点的选取更契合数据分布情况，减少算法迭代次数。通过计算数据集规模和数据间距离得到每个点不同的权值。d-K-means 算法基于“最小最大原则”选择聚类中心，避免了初始聚类中心选择随机性引起的局部最优解问题，更好地处理离群数据和低密度区域数据，同时通过权值及 BWP 指标[12]自动确定聚类中心个数。本文算法通过与四种算法在五个数据集上进行实验对比，验证了本文提出的算法能明显提高聚类效果质量以及聚类准确率。

1　相关算法

1.1　Canopy-Kmeans 算法

Canopy 聚类算法[13,14]是在 K-means 算法对数据进行预分类的算法，并且可以在人为无法确定聚类中心个数时通过 canopy 生成的大圆个数来近似设置。canopy 通过两个人为确定的阈值 $t1$ 和 $t2$ 对数据集进行处理，可以将混乱的数据分类成若干个有一定规则的数据堆。划分后的数据集分类效果如图 1 所示。该算法流程如下：

1）确定两个阈值 $t1$、$t2$（$t1>t2$）。

2）从数据集合中随机选出一个数据，计算这个数据到 canopy 的距离（如果当前没有 canopy，则该点直接作为 canopy 中心点）。

3）如果这个距离小于 $t1$，则给这个数据标上弱标记，将 $t1$ 加入这个 canopy 中（同时这个数据可以作为新的 canopy 来计算其他数据到这个点的距离）。

4）如果这个距离小于 $t2$，则给这个数据标上强标记，并将其中数据集合中删除，此时认为这个数据点距离该 canopy 已经足够近了，不需要形成新的 canopy。

5）重复 2~4 的过程，直至数据集合中没有数据。

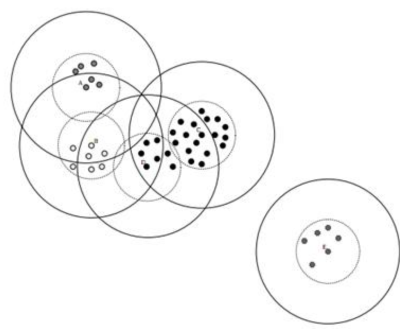


图 1　Canopy 分类效果图

Fig.1　Canopy classification effect

针对分类数量难以确定的问题，Canopy-Kmeans 算法可以通过 canopy 生成的大圆数量近似设置聚类中心个数，但算法实际应用时初始 Canopy 中心点和 Canopy 区域大小等初始值的选取好坏对聚类质量有较大影响。

1.2　K-means++算法

K-means++算法[15,16]按照如下的思想选取 K 个初始聚类中心：假设已经选取了 n 个初始聚类中心($0<n<K$)，则在选取第 $n+1$ 个聚类中心时距离当前 n 个聚类中心越远的点会有更大的概率被选为第 $n+1$ 个聚类中心。在选取第一个聚类中心($n=1$)时同样通过随机选取已有样本点的方法。可以说这也符合本文的直觉：聚类中心互相离得越远越好。这个改进虽然直观简单，却非常有效，很好地改进了 K-means 算法初始中心点选择[17]的随机性。概率计算函数为

$$P=\frac{D^2(x)}{\sum_{x\in X}D^2(x)}$$

其中：X 是聚类问题中点的集合；$D(x)$ 函数的定义为计算点到最近已选聚类中心的距离。观察概率函数发现，低密度区域的噪声点有较大可能被选为聚类中心，使属于此类中心点的数据过少，并且在之后的 K-means 算法迭代过程中发生改变的可能性很小，造成达不到 k 个聚类中心应有的分类效果。由两个簇构成的数据集,其中一个簇含有一个噪声点如图 2 所示。从图 3、4 的实验数据可以看到 K-means++算法对数据集的划分共有两种情况，值得注意的是即使不考虑噪声点，K-means++算法也无法得到正确的聚类结果[18]。

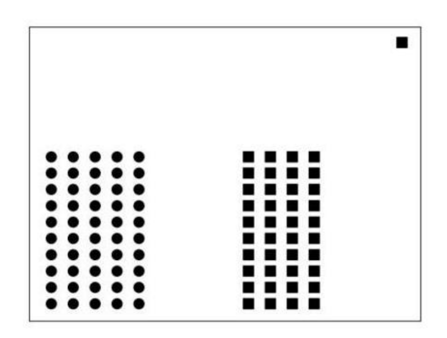


图 2　由两个簇构成的数据集,其中一个簇含有一个噪声点

Fig. 2　Data set composed of two clusters,one of the clusters contains a noise point

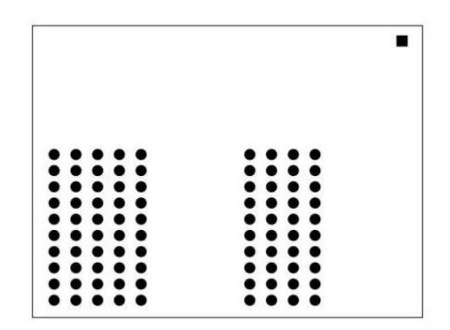


图 3　K-means++算法分类情况一

Fig.3　K-means++ algorithm classification case 1

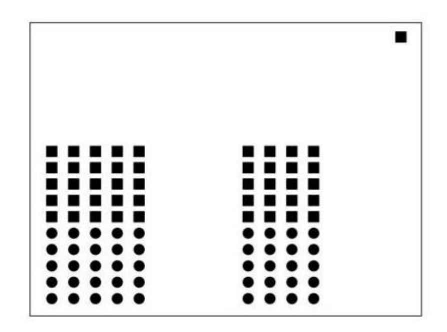


图 4　K-means++算法分类情况二

Fig.4　K-means++ algorithm classification case 2

1.3　DBSCAN 算法

DBSCAN 算法[19]是一种基于高密度连通区域的基于密度的聚类算法，可以自动确定簇的数量，具有噪声应用，通过找

出核心对象并与其邻域连接形成稠密区域作为簇。DBSCAN 的概念主要有：

定义 1　邻域。样本集合 D 中的一点 o，o 的邻域表示为以 o 为中心，ε 为半径的 d 维超球体区域。

定义 2　核心对象。一个对象的 ε-邻域至少包含 minPts 个对象。

定义 3　直接密度可达。样本集合 D 中，若对象 q 在核心对象 p 的 ε-邻域内，则 q 是从 p 的直接密度可达。

定义 4　密度可达。样本集合 D 中，若存在一个点链 p_1、p_2…、p_n，对于 $p_i \in D(1 \le i \le n)$，且 p_{i-1} 是从 p_i 直接密度可达，则点 p_n 从 p_1 密度可达。

定义 5　密度相连。若存在对象 o，使对象 p 和对象 q 都从 o 密度可达，则 p 和 q 密度相连。

DBSCAN 算法优点是能够基于密度分布发现任意形状的簇，可以很好地克服 K-means++的缺陷，但它的缺点[20]与大部分对 K-means 改进的算法相同：初始中心点的选择前提是点的 ε-邻域内数据点密度大，邻域内密度低的点会被标记为噪声，有可能剔除的正是用户所关心的数据，影响最终分类精度。同时，半径 ε 和最小支持度 minPts 的设定比较敏感。

2　d-K-means 算法

2.1　基本定义

d-K-means 算法中有关距离的计算均使用欧几里德距离，本文参照文献[19]依据贪心策略不断的自适应计算数据点 p 的半径，计算公式为

$$\varepsilon = \frac{1}{k}\sum_{p_i \in C} d(p_i, p_{k_nearest(i)})$$

其中：$p_{k_nearest(k)}$表示与点 p_i 最近的 k 个点；$d(\ )$表示点间的欧几里德距离；k 值在二维空间聚类中一般取 4[21]，其他情况可取数据集的$\lfloor n/25 \rfloor$[22]（其中：n 为数据样本总数；$\lfloor\rfloor$表示向下取整）。根据对象 p 与其 ε-邻域内对象 q 的距离计算 p 的权值，对权值进行加工得到每个数据的中心点指标 C_p，计算公式为

$$C_p = w_p * \theta_p$$

其中：w_p 为数据点 p 的权值，反映了点 p 的邻域密度；θ_p 为数据点 p 与距离自身最近的中心点 i 间的距离，计算公式分别为

$$\theta_p = \min_{1 \le i \le k} d(i, p)$$

$$w_p = \sum_{q=1}^{m} \frac{range - d(p,q)}{range}$$

其中：k 表示现有中心点个数；m 表示数据点 p 的 ε-邻域内数据对象数量；range 代表数据集向量空间的大小，计算方法与欧式距离计算方法相同，计算公式为

$$range = \sqrt{\sum_{z=0}^{x} \| \max_z - \min_z \|^2}$$

其中：x 表示数据集维度；max、min 代表数据集相应维度特征的最大值和最小值；$\| \|^2$ 表示欧氏距离的平方；$range$ 值实质上是数据集全部维度范围的模，点 p 的 ε-邻域内每个数据会为 w_p 贡献 0-1 的值，距离点 p 越近，贡献值越大。

此外，w_p 值越大，点 p 周围数据越多，数据越集中。θ_p 值越大，点 p 与已产生的聚类中心距离越远。w_p 与 θ_p 相乘得到的中心点指标 C_p 越大，簇内越紧密，两个簇之间的相异程度越高。K-means 算法时耗主要由迭代次数决定，通过中心点指标选择聚类中心可有效减少 K-means 算法迭代次数，提高算法时间性能。

d-K-means 算法参照文献[12]提出的一种新的聚类有效指标（称为 BWP 指标），根据 BWP 指标平均值的变化决定是否继续中心点的选取。BWP 指标平均值的计算公式为

$$\overline{BWP}(j,i) = \frac{1}{n}\sum_{i=1, i \in j}^{n} \frac{b(j,i) - w(j,i)}{b(j,i) + w(j,i)}$$

其中：n 为数据集规模大小；$b(j,i)$、$w(j,i)$定义如下：

存在 n 个数据对象的数据集 S，假设 n 个数据对象被划分到 k 个类中，定义第 j 类的对象 i 的类间距离 $b(j,i)$为该样本到其他每个类中样本平均值中的最小值，定义第 j 类的对象 i 的类内距离 $w(j,i)$为该数据对象到 j 类中其他数据对象距离的平均值，公式如下：

$$b(j,i) = \min_{1 \le c \le k, c \ne j}\left(\frac{1}{n}\right)\sum_{p=1}^{n_c} \|x_p^{(c)} - x_i^{(j)}\|^2$$

$$w(j,i) = \left(\frac{1}{n_j - 1}\right)\sum_{p=1, p \ne i}^{n_j} \| x_p^{(j)} - x_i^{(j)} \|^2$$

其中：c 和 j 表示类标；n_c 表示类 c 的元素个数；$x_p^{(c)}$表示第 c 类的第 p 个数据对象；$x_i^{(j)}$表示第 j 类的第 i 个对象。观察上式可知：$b(j,i)$越大类间分离性越好，$w(j,i)$越小类内紧密性越好，则 BWP 指标值越大聚类效果越好。

2.2　d-K-means 算法描述

d-K-means 算法根据如下方法依次选取中心点：基于“最小最大原则”，每次选取中心点指标值最大的数据点作为实验聚类中心进行预分类，即把所有数据点归为距离自身最近的中心点代表的类中，比较预分类前后所有数据点 BWP 指标平均值的变化。若 BWP 指标平均值变大，则此点成为聚类中心，参照 canopy 聚类算法思想，新聚类中心的 ε-邻域内数据点不参与接下来中心点的选取。同时，新聚类中心的产生可能会使数据点的最近中心点变化，因此每产生一个中心点便更新一次全部数据的中心点指标。若 BWP 指标平均值变小或不存在数据点可被选取则停止选取中心点，通过上述方法可以自动确定 k 个聚类中心。

d-K-means 算法中心点的选取思想实际聚类效果类似：聚类中心之间有一定距离，聚类中心周围的点密度较大。观察权值计算公式可以发现：数据点周围密度越大，数据点权值越大。将“最小最大原则”应用在中心点指标上，那么如果一个数据点权值较大，且距离类中心点距离较远，则此点被选为中心点的可能性较大。

聚类开始时无中心点，数据的中心点指标缺少 θ 参数无法计算。由于当数据对象在给定空间范围内数据点越多，说明该数据对象作为聚类中心店更有利于目标函数的收敛，所以选取权值最大的点作为第一个中心点，利于提高簇内紧密性，也符合文献[8,9]的中心点选取思想和实际聚类效果：中心点周围密度较大。中心点选取过程如图 5~8 所示。算法首先选择权值最大点作为第一个中心点，然后根据中心点指标依次选取了两个中心点。由于产生第四个中心点时 BWP 指标平均值减小，所以停止选取，得到前三个中心点。

d-K-means 算法通过权值和中心点指标的引入权衡了距离和密度对聚类的影响，中心点指标的引入让与当前聚类中心距离较远，但权值并不大的低密度区域数据也可能成为聚类中心。通过“最小最大原则”和第一个中心点的选择策略消除了初始中心点选择的随机性。通过 BWP 指标的引入与 canopy 算法思想的结合，算法可以自动确定类数；同时，通过 BWP 指标增减决定是否继续选择中心点，可以避免因结合 canopy 算法思想而在面对离群数据与噪声数据时将其选为单独一类的情况，保证了聚类效果的质量。

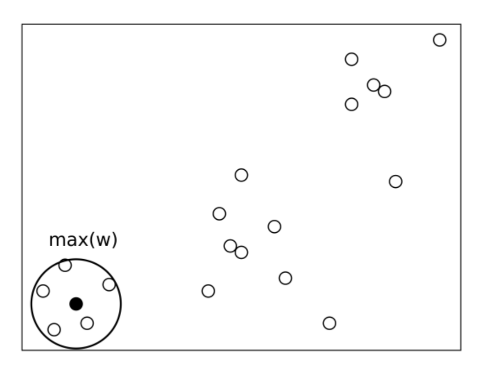


图 5　步骤一

Fig.5　Step1

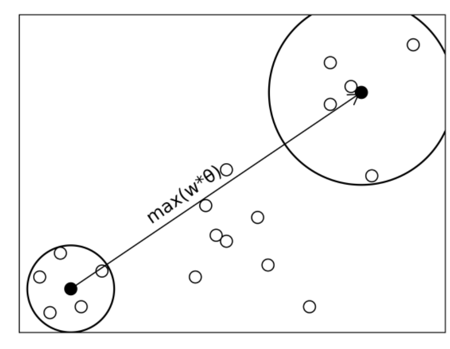


图 6　步骤二

Fig.6　Step2

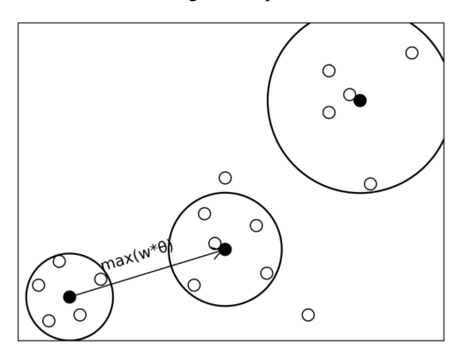


图 7　步骤三

Fig.7　Step3

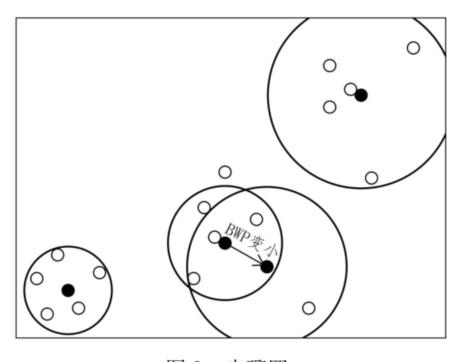


图 8　步骤四

Fig.8　Step4

2.3　d-K-means 算法流程

输入： n 个数据对象的集合 X。

输出： k 个聚类中心。

a）计算 n 个数据的 ε 半径和权值。

b）选择最大权值数据点作为第一个聚类中心。

c）计算每个数据点的中心点指标，选取中心点指标最大的数据点进行预分类。

d）计算预分类后 n 个数据点平均 BWP 指标值。

e）若 BWP 指标平均值增加，此数据点成为聚类中心，使其 ε-邻域内的数据不参与接下来中心点的选取，转向步骤6；若 BWP 指标平均值减小，转向步骤 g)。

f）若存在可以聚类的数据点，转向步骤 c)，否则转向步骤 g)。

g）将产生的中心点作为初始聚类中心执行 K-means 算法，得到最终聚类结果。

算法流程如图 9 所示。

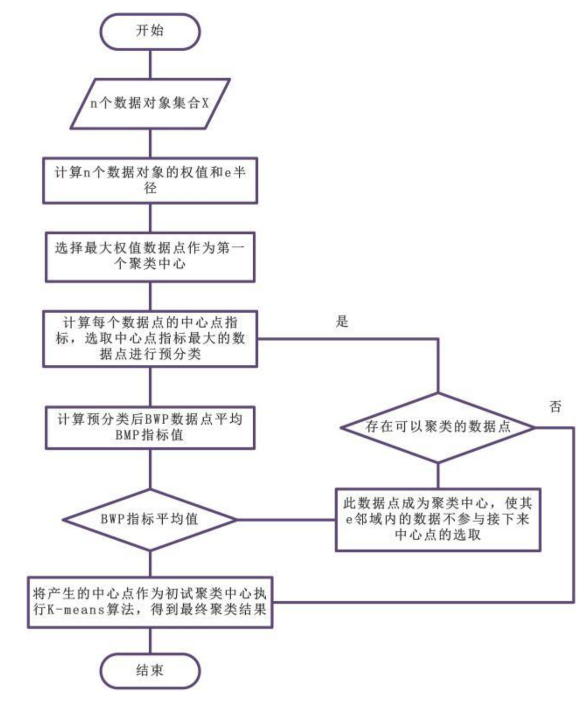


图 9　d-K-means 算法流程

Fig. 9　Flow chart of d-K-means algorithm

3　实验数据

3.1　数据集

为了验证 d-K-means 算法在选取初始聚类中心的有效性，选取了专门用于测试聚类算法性能的 UCI 数据集和美国气象局气候分类数据（Weather）。UCI 是加州大学提出的用于机器学习的数据库，是一个常用的标准测试数据集，实验数据集中都有明确的分类，因此可以直接观察到聚类的质量。实验对 Iris、Wine、Seeds、Pima、Weather 五个数据集进行测试，将 BWP 指标、兰德指数、轮廓系数、Jaccard 系数、迭代次数、准确率作为性能指标，与传统 K-means 算法、K-means++算法、文献[10]算法、文献[11]算法进行比较，表 1 为数据集参数情况。

表 1　数据集描述

Table 1　Description of data set

数据集	样本数	特征数	类数
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Pima	768	8	2
Weather	19	2	4

3.2　实验结果

1）在 UCI 数据集上的测试

数据集中不同的特征往往具有不同的量纲，这样的情况会影响到最终的分类效果。为了消除指标之间的量纲影响，对数据进行了归一化的预处理，使每个特征处于同一量级，合理地参与算法的执行。具体使用的是 Min-Max Scaling 方法。

对于拥有 i 个特征的数据集，对每个特征进行归一化处理[23]，公式如下：

$$z=\frac{x_i-\min(x_i)}{\max(x_i)-\min(x_i)}$$

图 10~15 分别为为本文算法与四种对比算法在 Iris、Wine、Seeds、Pima 数据集上对应的 BWP 指标、兰德指数、轮廓系数、Jaccard 系数、迭代次数、准确率。

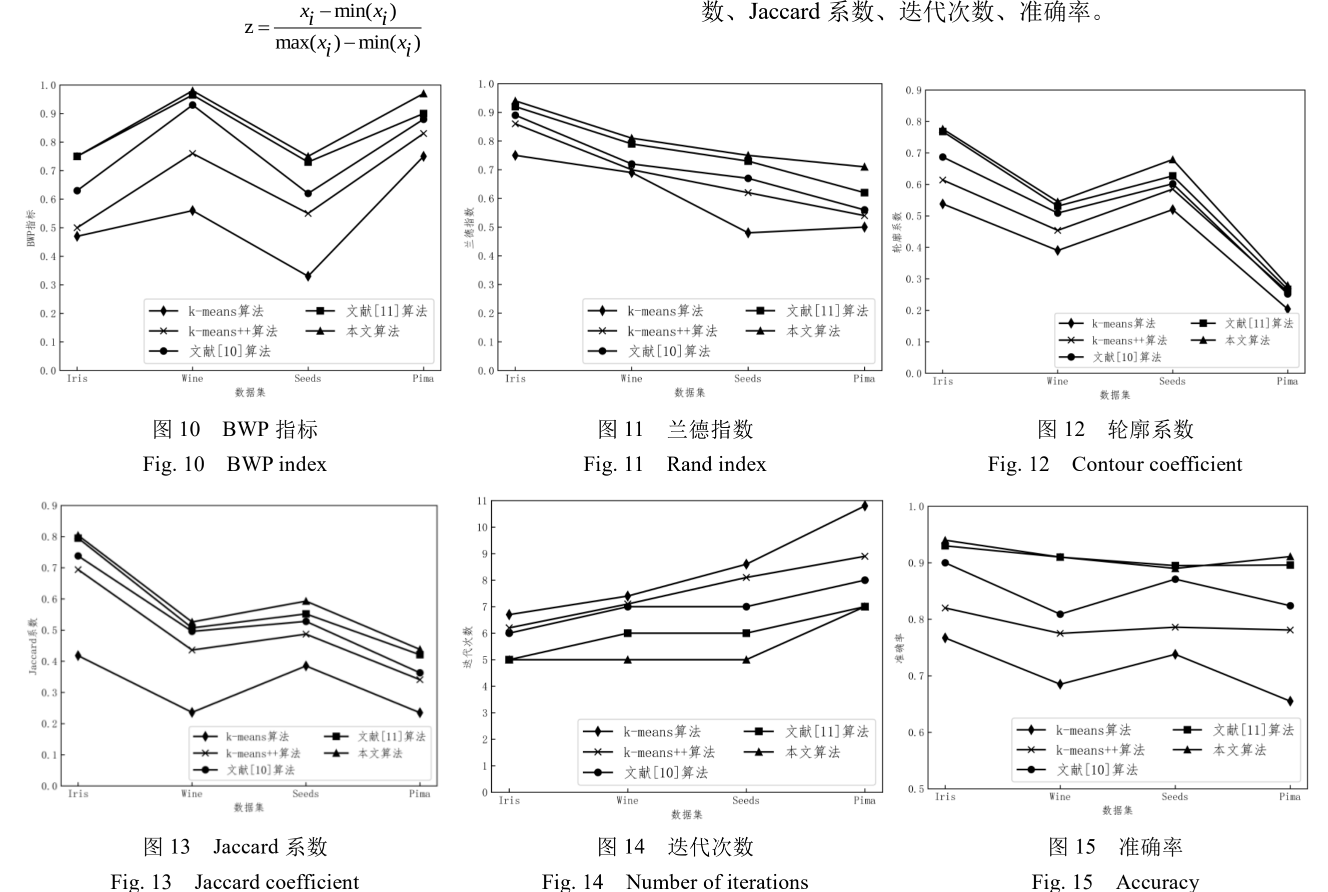


图 10　BWP 指标

Fig. 10　BWP index

图 11　兰德指数

Fig. 11　Rand index

图 12　轮廓系数

Fig. 12　Contour coefficient

图 13　Jaccard 系数

Fig. 13　Jaccard coefficient

图 14　迭代次数

Fig. 14　Number of iterations

图 15　准确率

Fig. 15　Accuracy

从图 10~13 的实验结果可以看出，d-K-means 算法得出的聚类结果的 BWP 指标、兰德指数、轮廓系数和 Jaccard 系数明显优于传统 K-means 算法、K-means++算法与文献[10]算法。原因是 K-means 算法随机选取初始中心点，K-means++算法中心点选取时虽然融入了距离，使与已产生的中心点距离远的数据更有可能被选为下一个聚类中心，但选择仍存在随机性，而文献[10]算法根据密度参数选择初始中心点，在计算密度参数时，没有考虑数据间的距离，密度参数计算不科学，影响初始聚类中心的质量。在 Iris、Wine、Seeds 数据集上，d-K-means 算法与文献[11]算法得出的聚类结果的 BWP 指标值与兰德指数值基本相同，在 Pima 数据集上，d-K-means 算法取得的聚类结果的 BWP 指标值与兰德指数值优于文献[11]算法，原因是 Pima 数据集有缺失值，而文献[11]算法易受离群点、噪声点影响，通过实验结果可以发现本文算法对样本的聚类效果较好，并且能更好地处理异常情况，鲁棒性较好。同时，本文算法在四个数据集上聚类结果的 BWP 指标、轮廓系数和 Jaccard 系数表现都是最好的，说明使用 d-K-means 算法可以得到更好的聚类效果：更高的类内紧密度与类间分离度。

在实验过程中发现，由于 K-means 算法与 K-means++算法初始聚类中心的选取都具有一定随机性，导致准确率和迭代次数不稳定，因此对两种算法进行 10 次实验，取实验结果的平均值进行对比。从图 14 与 15 可以看出，除了在 Seeds 数据集上本文算法准确率比文献[11]算法稍差以外，本文提出的 d-K-means 算法在其他数据集上迭代次数与准确率都是最优的，可见 d-K-means 算法处理低维数据和高维数据时都可以取得较好的聚类效果与效率。

2）在 Weather 数据集上的测试

Weather 数据集包括美国气象局 2018 年 9 月 12 日的交通要塞测量数据及所属类别，数据分布离散，低密度区域数据较多。实验根据对湿度和华氏温度两个特征对这些地点进行分类，相同的，对湿度和华氏温度进行了归一化处理。表 2 为 Weather 数据。图 16 为数据分布图。将 d-K-means 算法与 K-means 算法、K-means++算法、文献[10]算法、文献[11]算法在 Weather 数据集上运行，所得聚类结果及准确率如表 3 所示。

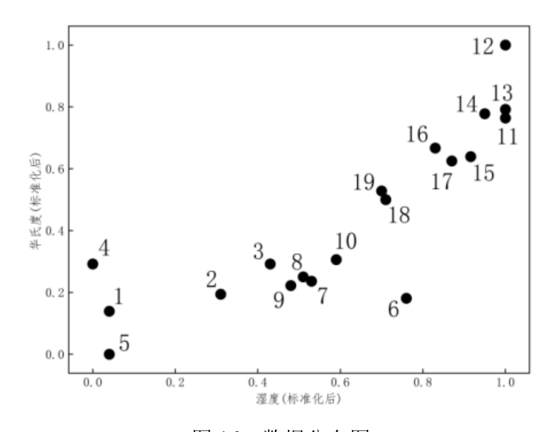


图 16　数据分布图

Fig. 16　Data distribution diagram

文献[10]算法与文献[11]算法结合了 canopy 思想，中心点的数据集平均距离范围内的数据不参与接下来的聚类，计算公式为

$$MeanDis(D)=\frac{2}{n(n-2)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}d(x_i,x_j)$$

由于 Weather 数据集的数据离散度较高，数据集平均距离较大，导致文献[10]算法与文献[11]算法得不到正确的分类数。本次实验中，K-means 算法通过随机选取确定初始聚类中心，

选取的点分别为(0.134,0.86)、(0.413,0.873)、(0.531,0.229)、0.912,0.875)。在随后的迭代过程中，没有任何数据属于前两个初始中心点，导致只生成两类，算法迭代三次，准确率为 52.6%。

K-means++算法在选取中心点时考虑了距离，距离已产生中心点远的数据更有可能被选为下一个聚类中心，第一个中心点的选取是随机的，剩余中心点的选取也存在随机性。依次选取数据 4、数据 11、数据 10、数据 6 为初始聚类中心，算法迭代三次，准确率为 57.9%。

文献[10]算法通过寻找数据点固定距离内的点确定密度参数，数据集平均距离为 0.523 5，依次选取数据 8、数据 16、数据 1 为初始聚类中心，选完上述三点后数据集中不存在点可以被选为聚类中心，因此生成三类，算法迭代两次，准确率为 68.4%。

表 2　美国地区气候数据统计

Table 2　Statistical table of regional climate data in the United States

标识	城市名称	湿度		华氏度/°F		类别
		原始	标准后	原始	标准后	
1	Caliente	20%	0.04	23	0.139	Red Flag Warning
2	Winnemucca	43%	0.31	27	0.194	Red Flag Warning
3	BRIGHAM	53%	0.43	34	0.292	Red Flag Warning
4	Sand Springs Smt	17%	0	34	0.292	Red Flag Warning
5	DEXTER	20%	0.04	13	0	Red Flag Warning
6	FORT ROCK	80%	0.76	26	0.181	Frost Advisory
7	Alturas	61%	0.53	30	0.236	Frost Advisory
8	Lake County Airport	59%	0.51	31	0.25	Frost Advisory
9	Oklahoma City	57%	0.48	29	0.222	Frost Advisory
10	Klamath Falls International	66%	0.59	35	0.306	Frost Advisory
11	Conway Horry County	100%	1	68	0.764	Hurricane Warning
12	Duplin County	100%	1.00	85	1	Hurricane Warning
13	Georgetown County	100%	1.00	70	0.792	Hurricane Warning
14	North Myrtle Beach Grand Strand	96%	0.95	69	0.778	Hurricane Warning
15	Marysville Municipal	93%	0.92	59	0.639	Hazardous Weather
16	Blosser Municipal	86%	0.83	61	0.667	Hazardous Weather
17	Newton City	89%	0.87	58	0.625	Hazardous Weather
18	Deep Creek US	76%	0.71	49	0.5	Hazardous Weather
19	RUBY VALLEY	75%	0.70	51	0.528	Hazardous Weather

表 3　算法运行结果比较

Table 3　Comparison of algorithm running results

算法	Hurricane Warning 地点标识	Hazardous Weather 地点标识	Frost Advisory 地点标识	Red Flag Warning 地点标识	迭代次数	准确率
K-means 算法	无	11,12,13,14,15,16,17,18,19	1,2,3,4,5,6,7,8,9,10	无	3	0.526
K-means++算法	11,12,13,14,15,16,17	6	2,3,7,8,9,10,18,19	1,4,5	3	0.579
文献[10]算法	无	11,12,13,14,15,16,17,18,19	2,3,6,7,8,9,10	1,4,5	2	0.684
文献[11]算法	无	11,12,13,14,15,16,17	2,3,6,7,8,9,10,18,19	1,4,5	2	0.684
本文算法	11,12,13,14	15,16,17,18,19	2,3,6,7,8,9,10	1,4,5	2	0.895

文献[11]算法与文献[10]算法思想类似，依次选取数据 8、数据 14、数据 1 为初始聚类中心，选完上述三点后数据集中不存在点可以被选为聚类中心，因此生成三类，算法迭代两次，准确率为 68.4%。

d-K-means 算法中每个数据的 ε 半径都是自适应计算的，因此使中心点的 ε-邻域数据不参与进一步的聚类更加科学，并且 d-K-means 算法中数据权值的计算结合了数据间的距离，相较于文献[10]提出的密度参数，权值可以更好地反映密度分布情况，在权值基础上计算得到中心点指标，通过中心点指标选取聚类中心让聚类权衡了距离与密度的关系。算法依次选取数据 8、数据 13、数据 1、数据 19 为初始聚类中心，算法迭代两次，准确率为 89.5%。

通过算法在五个数据集上的实验结果发现：无论应对低维数据或高维数据，离散数据或密集数据，d-K-means 算法都可以取得较好的聚类效果、较高的聚类精度与迭代速度。

4　结束语

随着如今数据的迅速膨胀与变大，数据分布情况更加多样化，为了节约时间，自动选择聚类中心个数的需求也越来越大。d-K-means 算法通过权值与中心点指标的引入权衡了距离与密度的关系，相对于 K-means 算法容易受初始点选择的影响陷入局部最优以及需要手动选择中心点个数的缺陷，基于距离和密度的 d-K-means 算法不仅可以自动选择中心点个数，并且聚类效果更好，在迭代速度以及分类准确性上都有所提高。

参考文献：

[1] Han Jiawei, Kamber M, Pei Jian, 等. 数据挖掘：概念与技术 [M]. 范明，孟小峰，译. 3 版. 北京：机械工业出版社, 2012: 211-213.（Han Jiawei, Kamber M, Pei Jian, *et al*. Data mining: concept and technology [M]. Fan Ming, Meng Xiaofeng, Translated. 3 edi. Beijing: Machinery Industry Press, 2012: 211-213.）

[2] Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods for the K-means clustering algorithm [J]. Expert Systems with Applications, 2013, 40 (1): 200-210.

[3] Bagirov A M. Modified global K-means algorithm for minimum sum-of-squares clustering problems [J]. Pattern Recognition, 2008, 41

(10): 3192-3199

[4] Tzortzis G, Likas A. The MinMax K-means clustering algorithm [J]. Pattern Recognition, 2014, 47 (7): 2505-2516

[5] 蔡龙飞. 运用硬C-均值改进K-means算法的聚类分析 [J]. 科技咨询导报，2007 (24): 144-145. (Cai Longfei. Clustering analysis of improved K-means algorithm using hard C-means [J]. Science and Technology Innovation Herald, 2007 (24): 144-145.)

[6] 冯波，郝文宁，陈刚，等. K-means 算法初始聚类中心选择的优化 [J]. 计算机工程与应用，2013, 49 (14): 182-185. (Feng Bo, Hao Wenning, Chen Gang, *et al.* Optimization to K-means initial cluster centers[J]. Computer Engineering and Applications, 2013, 49 (14): 182-185.)

[7] 翟东海，鱼江，高飞，等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究 [J]. 计算机应用研究，2014, 31 (3): 713-715,719. (Zhai Donghai, Yu Jiang, Gao Fei, *et al.* K-means text clustering algorithm based on initial cluster centers selection according to maximum distance [J]. Application Research of Computers, 2014, 31 (3): 713-715,719.)

[8] 唐东凯，王红梅，胡明，等. 优化初始聚类中心的改进 K-means 算法 [J]. 小型微型计算机系统，2018, 39 (8): 1819-1823. (Tang Dongkai, Wang Hongmei, Hu Ming, *et al.* Optimizing initial cluster center of improved K-means algorithm [J]. Journal of Chinese Computer Systems, 2018, 39 (8): 1819-1823.)

[9] 周炜奔，石跃祥. 基于密度的K-means聚类中心选取的优化算法 [J]. 计算机应用研究，2012, 29 (5): 1726-1728. (Zhou Weiben Shi Yuexiang. Optimization algorithm of K-means clustering center of selection based on density [J]. Application Research of Computers, 2012, 29 (5): 1726-1728.)

[10] 贾瑞玉，宋建林. 基于聚类中心优化的 K-means 最佳聚类数确定方法 [J]. 微电子学与计算机，2016, 33 (5): 62-66,71. (Jia Ruiyu, Song Jianlin. K-means optimal clustering number determination method based on clustering center optimization [J]. Microelectronics & Computer, 2016, 33 (5): 62-66,71.)

[11] Zhang Geng, Zhang Chengchang, Zhang Huayu. Improved K-means algorithm based on density canopy [J]. Knowledge-Based Systems, 2018, 145: 289-297.

[12] 王法胜，鲁明羽，赵清杰，等. 粒子滤波算法 [J]. 计算机学报，2014, 37 (8): 1679-1694. (Wang Fasheng Lu Mingyu Zhao Qingjie, *et al.* Particle filtering algorithm [J]. Chinese Journal of Computers, 2014, 37 (8): 1679-1694.)

[13] 张琳，牟向伟. 基于 Canopy+K-means 的中文文本聚类算法 [J]. 图书馆论坛，2018, 38 (6): 113-119. (Zhang Lin, Mou Xiangwei. Chinese text clustering algorithm based on canopy+K-means [J]. Library Tribune, 2018, 38 (6): 113-119.)

[14] 毛典辉. 基于 MapReduce 的 Canopy-Kmeans 改进算法 [J]. 计算机工程与应用，2012, 48 (27): 22-26. (Mao Dianhui. Improved canopy-Kmeans algorithm based on MapReduce [J]. Computer Engineering and Applications, 2012, 48 (27): 22-26.)

[15] Yoder J, Priebe C E. Semi-supervised K-means+ [J]. Journalof Statistical Computation & Simulation, 2016 (3)

[16] 张亚洲，余正生. 基于 K-means+聚类的视频摘要生成算法 [J]. 工业控制计算机, 2017, 30 (7): 129-130. (Zhang Yazhou, Yu Zhengsheng. Video summarization generation algorithm based on K-means+clustering [J]. Industrial Control Computer, 2017, 30 (7): 129-130.)

[17] Brunsch T, Röglin H. A bad instance for K-means+ [J]. Theoretical Computer Science, 2013, 505 (9): 19-26..

[18] Agarwala M, Jaiswalb R, Pal A. K-means+under approximation stability [J]. Theoretical Computer Science, 2015, 588: 37-51.

[19] 冯振华，钱雪忠，赵娜娜. Greedy DBSCAN：一种针对多密度聚类的 DBSCAN 改进算法 [J]. 计算机应用研究，2016, 33 (9): 2693-2696,2700. (Feng Zhenhua, Qian Xuezhong, Zhao Nana. Greedy DBSCAN: an improved DBSCAN algorithm on multi-density clustering [J]. Application Research of Computers, 2016, 33 (9): 2693-2696,2700.)

[20] Nasibov E N; Ulutagay G. Robustness of density-based clustering methods with various neighborhood relations [J]. Fuzzy Sets And Systems, 2009, 160 (24): 3601-3615

[21] 孙凌燕. 基于密度的聚类算法应用研究 [D]. 太原：中北大学, 2009. (Sun Lingyan. Research of clustering algorithm based on density [D]. Taiyuan: North University of China, 2009.)

[22] Daszykowski M, Walczak B, Massart D L. Looking for natural patternsin data: part 1. density-based approach [J]. Chemometrics andIntelligent Laboratory Systems, 2001, 56 (2): 83-92.

[23] 汤荣志，段会川，孙海涛. SVM 训练数据归一化研究 [J]. 山东师范大学学报 :自然科学版 , 2016, 31 (4): 60-65. (Tang Rongzhi, Duan Huichuan Sun Haitao. Research on data normalization for SVM training [J]. Journal of Shandong Normal University :Natural Science , 2016, 31 (4): 60-65.)